Atmospheric Environment 45 (2011) 6938–6946

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Impact of source collinearity in simulated $PM_{2.5}$ data on the PMF receptor model solution

Rima Habre[a,*], Brent Coull[b], Petros Koutrakis[c]

[a] Department of Environmental Health, Harvard School of Public Health, 401 Park Drive, Landmark Center West, Room 412-E, Boston, MA 02115, USA
[b] Department of Biostatistics, Department of Environmental Health, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
[c] Department of Environmental Health, Harvard School of Public Health, 401 Park Drive, Boston, MA 02115, USA

ARTICLE INFO

Article history:
Received 10 June 2011
Received in revised form
12 September 2011
Accepted 16 September 2011

Keywords:
PMF
Receptor model
Source collinearity
Simulation
$PM_{2.5}$
Source apportionment

ABSTRACT

Positive Matrix Factorization (PMF) is a factor analytic model used to identify particle sources and to estimate their contributions to $PM_{2.5}$ concentrations observed at receptor sites. Collinearity in source contributions due to meteorological conditions introduces uncertainty in the PMF solution. We simulated datasets of speciated $PM_{2.5}$ concentrations associated with three ambient particle sources: “Motor Vehicle” (MV), “Sodium Chloride” (NaCl), and “Sulfur” (S), and we varied the correlation structure between their mass contributions to simulate collinearity. We analyzed the datasets in PMF using the ME-2 multilinear engine. The Pearson correlation coefficients between the simulated and PMF-predicted source contributions and profiles are denoted by “*G* correlation” and “*F* correlation”, respectively. In sensitivity analyses, we examined how the means or variances of the source contributions affected the stability of the PMF solution with collinearity. The % errors in predicting the average source contributions were 23, 80 and 23% for MV, NaCl, and S, respectively. On average, the NaCl contribution was overestimated, while MV and S contributions were underestimated. The ability of PMF to predict the contributions and profiles of the three sources deteriorated significantly as collinearity in their contributions increased. When the mean of NaCl or variance of NaCl and MV source contributions was increased, the deterioration in *G* correlation with increasing collinearity became less significant, and the ability of PMF to predict the NaCl and MV loading profiles improved. When the three factor profiles were simulated to share more elements, the decrease in *G* and *F* correlations became non-significant. Our findings agree with previous simulation studies reporting that correlated sources are predicted with higher error and bias. Consequently, the power to detect significant concentration-response estimates in health effect analyses weakens.

1. Background

Exposure to ambient fine particles of aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) has been shown to be associated with significant increases in cardiovascular and respiratory morbidity and mortality risk (Dockery et al., 1993; Pope et al., 2002). Consequently, research interest has grown in investigating the differential toxicities of specific $PM_{2.5}$ components and sources, in order to inform air quality management decisions.

Receptor models rely upon $PM_{2.5}$ chemical speciation data to identify particle sources and to estimate their contributions to concentrations observed at a receptor site. The two main classes of receptor models, described in Hopke et al. (2006) and Miller et al. (2002), are: i) Chemical Mass Balance (CMB), where the knowledge of the number and chemical profile of sources is assumed; and ii) Factor analysis methods, where both the number and chemical profile of sources are unknown. The latter class of receptor models includes Principal Component Analysis (PCA) and its variations, UNMIX, and Positive Matrix Factorization (PMF). These models estimate the number of “factors” that explain the largest amount of variability or variance in the observed data. In addition, they decompose a matrix of chemical composition data into a contributions matrix, a chemical profiles matrix, and residual error. It is important to pursue systematic investigations about potential sources of uncertainty and bias, especially if source apportionment results are used for further analysis. For instance, epidemiological studies use estimated source contributions to link health endpoints to different source types (Laden et al., 2000; Ito et al., 2006). While the consistency of source apportionment results across several

* Corresponding author. Tel.: +1 617 384 8837; fax: +1 617 384 8859.
E-mail addresses: rhabre@hsph.harvard.edu, rimahabre@gmail.com (R. Habre), bcoull@hsph.harvard.edu (B. Coull), petros@hsph.harvard.edu (P. Koutrakis).

1352-2310/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2011.09.034

users and methods and the robustness of applications to $PM_{2.5}$ health effect assessments has been established (Thurston et al., 2005; Hopke et al., 2006; Ito et al., 2006), issues regarding the accuracy and uncertainty of their results still remain (Ito et al., 2004, 2006; Grahame and Hidy, 2007).

The Positive Matrix Factorization (PMF) model is used extensively in air pollution source apportionment studies (Lee et al., 1999, 2006; Larsen and Baker, 2003; Kim and Hopke, 2007). PMF is an individually weighted factor analytic model with non-negativity constraints, initially developed by Paatero and Tapper (Paatero and Tapper, 1993, 1994) and further refined by Paatero and Hopke (Paatero and Hopke, 2003). PMF is made available by the United States Environmental Protection Agency under a public license (Norris et al., 2008) or by the author Dr. Pentii Paatero under an individual license.

Uncertainty and bias in the PMF solution can potentially result from rotational ambiguity, sampling and measurement error, analytical uncertainties in the quantification of trace elements, day-to-day variability in source profiles and collinearity in source contributions (Paatero et al., 2005; Reff et al., 2007). Collinearity may stem from the variability of meteorological conditions that govern the dispersion and transformation of pollutants, as evidenced by high correlations among the different $PM_{2.5}$ constituents (Grahame and Hidy, 2007; Gent et al., 2009; Hemann et al., 2009).

Several studies have used simulated data to evaluate the sensitivity of the PMF model fit to different sources of uncertainty (Miller et al., 2002; Brinkman et al., 2006; Bzdusek and Christensen, 2006; Christensen and Schauer, 2008; Hemann et al., 2009). For example, Christensen and Schauer (2008) used data on $PM_{2.5}$ carbon fractions, ions, and metals from the St. Louis-Midwest Supersite and simulated three cases of increasing perturbation in the uncertainty matrix to evaluate the stability of the PMF model solution. They found that the relative errors increased across three scenarios representing increasing levels of uncertainty. Moreover, they reported that errors associated with the estimation of daily source contributions can be more than double those associated with the estimation of average source contributions. The stability of source profile estimates varied across sources, with secondary sulfate and secondary nitrate sources having the most stable source contribution estimates. Brinkman et al. (2006) simulated nine speciated $PM_{2.5}$ and organics personal exposure datasets to evaluate the accuracy of the PMF model under different filter chemical analysis scenarios, incorporating source profile variability and measurement uncertainty in the simulated concentrations. They found similar errors in source apportionment for all scenarios, where factors with uniform contributions or factors lacking unique tracers with concentrations above detection limit were harder to separate. However, they noted that when contributions from a pair of sources were highly correlated in the simulated data, PMF usually resolved a single factor corresponding to both sources.

Hemann et al. (2009) simulated one year of daily ambient speciated $PM_{2.5}$ data, comprised of 39 species of carbon fractions, ions, and organics from nine pollutant sources. Toward this end, they used neural networks and bootstrapping methods to estimate the bias and variability due to random sampling error in predicted source contributions at the daily time scale. They found that the three factors, “Gasoline Vehicles”, “Meat Cooking”, and “Natural Gas”, that had moderately strong correlations ranging from 0.64 to 0.81 in their simulated contributions time series, had the highest bias and variability in their solutions. They also found that while PMF was able to fit the contributions of some sources reasonably well, bias in the solution can be high even when the variability or uncertainty is low. Finally, Miller et al. (2002) used Monte Carlo techniques to simulate personal exposure to 13 volatile organic compounds and compared the performance of PMF to CMB, UNMIX, and PCA/Abs Principal Component models. They found that PMF was best able to predict the major input factor profiles; however, all four receptor models were unable to properly separate sources whose contributions are strongly correlated or profiles are similar.

While investigating different sources of error and bias in the PMF solution, these simulation studies have reported that correlated sources are generally harder to resolve, and factors such as source variability, mass concentrations, or uniqueness of the source profiles seem important. However, to the best of our knowledge, the impact of source contributions collinearities on the PMF receptor model fit has not been directly and systematically investigated yet, and the relative importance of such factors is not well understood. Therefore, the aim of our study is to simulate increasing collinearity among source contributions and to examine the impact of such correlation on the ability of PMF to estimate source profiles and contributions. Furthermore, using sensitivity analyses, we examine whether changes in the means or variances of the source contributions, or similarity of the source profiles, affect the stability of the PMF solution with the presence of collinearity.

2. Methods

2.1. Data simulation

Datasets of speciated $PM_{2.5}$ concentrations (C) and their respective uncertainties (U) are simulated using the R v2.9.1 statistical package (R Development Core Team, 2009). Each simulated dataset consists of 340 daily concentrations of $PM_{2.5}$ mass and 18 elements that are associated with three ambient particle sources: “Motor Vehicle” (MV), “Sodium Chloride” (NaCl), and “Sulfur” (S).

The concentration matrix ($C(m\times n)$) is defined as

$$C(m \times n) = G(m \times p)^{*}F(p \times n) + E(m \times n) \tag{1}$$

where $G(m\times p)$ is the source contribution matrix (m is the number of daily samples, equal to 340 and p is the number of sources, equal to 3); $F(p\times n)$ is the source profile matrix (n is the number of elements/species, equal to 18); and $E(m\times n)$ is the error matrix that includes both analytical and measurement errors. Therefore, the concentration of an element j on day i can be written as follows:

$$C_{ij} = G_{ik}{}^{*}F_{jk} + e_{ij} \tag{2}$$

where k indexes source.

In the primary analysis, the Motor Vehicle, Sodium Chloride, and Sulfur mass contributions in the “G” matrix are each simulated from a normal distribution, with means of 16, 6, and 20 μg m^{-3}, respectively, and a covariance matrix “Sigma”, where Sigma = Var * Corr. “Var” is the variance matrix with 21, 6, and 45 respectively on the diagonal, and “Corr” is the correlation matrix between the 3 sources.

Source collinearity is simulated by changing the correlation structure between the MV, NaCl, and S mass contributions in the “Corr” matrix. Twenty-seven combinations of collinearity scenarios are obtained by varying each pairwise correlation in “Corr” to take the value of 0.0, 0.3, or 0.6, resulting in twenty-seven correlation scenarios. For each scenario, 100 datasets are generated, resulting in 2700 datasets.

A secondary set of sensitivity analyses is aimed at examining how changes in the means or variances of the source contributions, or similarity of the source profiles, in the presence of increasing collinearity, affects the stability of the PMF model solution. To accomplish this, the means and variances of the three source contributions are individually varied. In order to capture the relative variance of a factor with respect to its mean, the coefficient of variation (CV) is defined as the standard deviation of the source contribution divided by its mean.

In each of the seven sensitivity analyses, only one parameter is varied in comparison to the primary analysis: 1) the mean of NaCl is increased keeping the original CV; 2) the variance of NaCl is increased; 3) the variance of NaCl is decreased; 4) the variance of MV is increased; 5) the CV's of all three factors are increased to the same level keeping the original means; 6) the means of all three factors are adjusted to 16 μg m^{-3} keeping the original CV's; and 7) the three factor profiles are made more similar. Each scenario in the sensitivity analyses is simulated 10 times (270 datasets each). Table 1 presents the means, variances, and coefficients of variation that were specified in the primary analysis and the secondary sensitivity analyses, along with the number of simulations. The PM$_{2.5}$ mass on day i is equal to

$$\sum_{k=1}^{3} G_{ik} \tag{3}$$

The "F" matrix of factor loading profiles encompassing 18 elements is adapted from the Gent et al. (2009) source apportionment study on speciated ambient PM$_{2.5}$ data collected between August 2000 and February 2004 in New Haven, Connecticut. In order to create similar profiles for the seventh sensitivity analysis, the loading of any element that was present in two of the three factors was set to the minimum of the two loadings in the third factor. This resulted in factor profiles that share 10 of the 18 elements in common, with 8 elements remaining unique to a single factor. The loading profiles used in the different simulations are presented in Table 2.

The matrix of analytical and measurement errors E is calculated as

$$E_{ij} = Z_{ij}{*}U_{ij} \tag{4}$$

where "Z" is a matrix of Z scores sampled from a Normal distribution for each element on each day ($Z_j \sim \mathrm{N}(0,1)$). "U" is the matrix of analytical uncertainties in the daily elemental concentrations. The uncertainty of element j on day i consists of 2 components: a fraction (f) of the concentration of element j on day i, and the Method Detection Limit (MDL) of element j. Therefore, the uncertainty of element j on day i is given by the following equation:

$$U_{ij} = \sqrt{\left(f{*}G_{ik}{*}F_{kj}\right)^2 + \mathrm{MDL}^2} \tag{5}$$

where $f = 0.2$ for PM$_{2.5}$ mass and $f = 0.1$ for elements.

Table 3 lists the MDL values used to simulate the daily uncertainties. For Elemental Carbon (EC) the MDL of 80 ng m^{-3} is used, which is based on the NIOSH 5040 thermal optical transmittance method (Dutton et al., 2009). For PM$_{2.5}$ mass, the typical MDL of FRM gravimetric analysis (0.3 μg m^{-3}) is used, and for the remaining 17 elements, the X-ray Fluorescence MDL is used. In order to decrease the weight of PM$_{2.5}$ mass in the PMF solution, its MDL is multiplied by 10 to a final value of 3 μg m^{-3}. Finally, in all data simulations, contributions and concentrations that were negative were set to zero.

Table 2
Factor loading profiles (ng μg^{-1}) used in the simulations. The primary analysis factor loading profiles, adapted from Gent et al. (2009), are used in all simulations except the "Similar Profiles" sensitivity analysis.

Element	Primary Analysis			"Similar Profiles" Analysis		
	Motor Vehicle	Sodium Chloride	Sulfur	Motor Vehicle	Sodium Chloride	Sulfur
EC	145.7	0.0	21.6	145.7	20.0	21.6
Zn	3.4	0.0	0.1	3.4	0.1	0.1
Pb	0.5	0.0	0.1	0.5	0.1	0.1
Cu	0.3	0.0	0.1	0.3	0.1	0.1
Se	0.1	0.0	0.0	0.1	0.0	0.0
Si	0.0	0.0	1.0	0.0	0.0	1.0
Fe	12.8	0.0	0.5	12.8	0.5	0.5
Al	0.0	0.0	1.3	0.0	0.0	1.3
Ca	2.7	0.0	0.0	2.7	0.0	0.0
Ba	0.4	0.0	0.1	0.4	0.1	0.1
Ti	0.2	0.0	0.1	0.2	0.1	0.1
S	16.7	25.0	164.9	16.7	25.0	164.9
P	2.1	0.0	6.7	2.1	2.1	6.7
K	0.0	0.0	2.3	0.0	0.0	2.3
V	0.1	0.0	0.1	0.1	0.1	0.1
Ni	0.3	0.0	0.0	0.3	0.0	0.0
Na	0.0	200.0	0.0	0.0	200.0	0.0
Cl	0.0	300.0	0.0	0.0	300.0	0.0

2.2. *Positive matrix factorization*

In each scenario, the sets of concentration and sample-specific uncertainty matrices are analyzed in PMF using the ME-2 multilinear engine executable (me2wopt.exe) and script file (PMF2_bs2.ini), under the EPA public license available with the EPA PMF v3.0 installation (Paatero, 1999, 2010; Norris et al., 2009). PMF uses an iterative least squares algorithm to solve equation (2) by minimizing the sum of squares object function Q for a given number of factors p. Q is defined as the following:

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{U_{ij}}\right)^2 \tag{6}$$

where e_{ij} is the residual error in the PMF model for species j on day i, and U_{ij} is the uncertainty of species j on day i, as defined in

Table 1
Simulation parameters used in the primary analysis and the seven sensitivity analysis scenarios.

Scenario	N Simulations	Simulation Parameters								
		Motor Vehicle			Sodium Chloride			Sulfur		
		Mean	Variance	CV	Mean	Variance	CV	Mean	Variance	CV
Primary Analysis	100	16	21	0.29	6	6	0.41	20	45	0.34
Sensitivity Analyses										
1) Increase NaCl Mean	10	16	21	0.29	10	16	0.40	20	45	0.34
2) Increase NaCl Variance	10	16	21	0.29	6	16	0.67	20	45	0.34
3) Decrease NaCl Variance	10	16	21	0.29	6	3	0.29	20	45	0.34
4) Increase MV Variance	10	16	40	0.40	6	6	0.41	20	45	0.34
5) Same Coefficients of Variation	10	16	41	0.40	6	6	0.41	20	64	0.40
6) Same Means	10	16	21	0.29	16	43	0.41	16	29	0.34
7) Similar Profiles	10	16	21	0.29	6	6	0.41	20	45	0.34

Table 3
Distribution of elemental concentrations and percent below detection limit in the primary analysis.

Element	Distribution		Detection Limit	
	Mean (ng m^{-3})	Std Dev (ng m^{-3})	MDL (ng m^{-3})	% < MDL
EC	2763	784	80	0
Zn	56	17	2	0
Pb	10	7	7	35
Cu	7	3	2	4
Se	2	1	1	30
Si	20	12	11	25
Fe	215	66	15	0
Al	27	17	15	27
Ca	43	13	2	0
Ba	18	22	33	77
Ti	5	2	2	9
S	3715	1215	3	0
P	168	52	4	0
K	46	17	5	1
V	4	1	1	3
Ni	5	3	3	30
Na	1201	510	87	1
Cl	1801	755	2	1

equation (5). Samples with high uncertainties will be down-weighted in the solution. Predicted profiles and contributions are constrained to being positive. The objective is to find the optimal values of G_{ik} and F_{kj} for a given number of sources p that fit C_{ij} with minimum Q value. The Q value is a goodness-of-fit parameter that assesses how the model fits the data. Q_{true} is calculated using all data points while Q_{robust} is calculated accounting for outlier points (points with scaled residuals above 4). The theoretical Q for a model run is equal to $nm - p(n + m)$, where n is the number of species, m is the number of samples, and p is the number of factors. In our simulations, $n = 19$ species (18 elements and $PM_{2.5}$ mass), $m = 340$ days, and $p = 3$ factors in each scenario, so $Q_{\text{theoretical}} = 5383$. Solutions with Q values within the range of $Q_{\text{theoretical}}$ are generally considered acceptable (Norris et al., 2008).

The PMF model is run in robust mode for all scenarios, with 10 base runs, 3 factors ($np = 3$), random seeds (contrun = 0), and 10% modeling uncertainty, in addition to the sample-specific uncertainties (error model code em = −12, $C1 = 0.0$, $C2 = 0.0$, $C3 = 0.1$). $PM_{2.5}$ mass is included in the factor analysis with high uncertainty, along with the 18 elements, such that the mass apportioned to each factor can be used to scale the normalized contributions to mass units, as suggested in Norris et al. (2009). There are no missing data, and no inclusion or deletion criteria are specified to ensure similar model runs across scenarios. The convergent PMF base run with the minimum Q_{robust} value and the least number of steps is selected as the model solution, and no FPEAK rotation or bootstrapping is performed on the base solution. Output files are generated by PMF and analyzed using the SAS 9.2 statistical package (SAS Institute Inc., 2011).

2.3. Data analysis

For the primary analysis, the means and standard deviations of the elemental concentrations, as well as the percent below MDL are calculated and presented in Table 3. In order to assign a predicted factor to a simulated factor in each PMF run, the Pearson correlation coefficient is calculated between the simulated (input) and predicted (output) source contributions (G correlation), as well as the simulated and predicted source profiles (F correlation). A predicted factor must have a Pearson correlation greater than or equal to 0.8 with the simulated contributions to be considered the estimate for that contribution. This test is performed on the contributions first, and if the G correlation is less than 0.8, then the correlation between the true and estimated profiles (F correlation) is examined. In addition to the G and F correlations, the percent absolute errors in estimating the daily and average mass contributions are calculated for each source as follows:

$$\%\text{Error}_{\text{Daily}} = \frac{\sum_{i=1}^{340} \frac{|g_{ik} - \hat{g}_{ik}|}{g_{ik}}}{340} * 100 \tag{7}$$

where g_{ik} and $\hat{g}_{ik}$ are the simulated and predicted contributions of source k on day i, respectively, and

$$\%\text{Error}_{\text{Average}} = \frac{|\bar{g}_k - \bar{\hat{g}}_k|}{\bar{g}_k} * 100 \tag{8}$$

where $\bar{g}_k$ and $\bar{\hat{g}}_k$ are the simulated and predicted average mass contributions of source k, respectively.

The degree of collinearity that is impacting each factor in each scenario is defined as the sum of the correlation imposed between that factor and the other two factors, and it ranges from 0 to 1.2. Therefore, for “Motor Vehicle”, “Sodium Chloride”, and “Sulfur”, the measure of collinearity is defined as $\text{Collinearity}_{\text{MV}} = \text{Corr}_{\text{MV,NaCl}} + \text{Corr}_{\text{MV,S}}$, $\text{Collinearity}_{\text{NaCl}} = \text{Corr}_{\text{NaCl,MV}} + \text{Corr}_{\text{NaCl,S}}$, and $\text{Collinearity}_{\text{S}} = \text{Corr}_{\text{S,MV}} + \text{Corr}_{\text{S,NaCl}}$, respectively.

In order to test the degree and significance of the change in the PMF prediction error against increasing collinearity, the % Daily Error, % Average Error, G correlation, and F correlation are separately regressed on collinearity for each factor in each of the simulation scenarios. For example, to test the magnitude and significance of the change in the G correlation for MV as collinearity is increased in the primary analysis, the slope and the p-value of the slope of the following linear regression are used: $G\ \text{Correlation}_{\text{MV}} = \text{Intercept} + \text{Slope} * \text{Collinearity}_{\text{MV}} + \text{Error}$. The change in G correlation is deemed significant if the p-value of the slope is less than or equal to 0.05.

3. Results

3.1. Primary analysis

Table 3 summarizes the distribution of the simulated elemental concentrations across the 100 simulations of 27 collinearity scenarios in the primary analysis. The elements Pb, Se, Si, Al, Ba, Ti and Ni have the highest percent below MDL, ranging between 9 and 77%.

All simulation scenarios converge in PMF, with Q_{robust} goodness-of-fit parameter values ranging from a minimum of a 2365 to a maximum of 2686 (mean 2512) in the primary analysis of 2700 datasets. With mean input contributions of 16, 6 and 20 μg m^{-3}, the mean predicted factor contributions of “Motor Vehicle”, “Sodium Chloride”, and “Sulfur” are 12.7, 10.8 and 15.8 μg m^{-3}, respectively. The distribution of the mean predicted contributions of the three sources in the primary analysis across all collinearity scenarios is shown in Fig. 1.

The % error for the estimation of daily contributions of “Motor Vehicle”, “Sodium Chloride”, and “Sulfur” is 27, 83, and 27%, respectively. The increase in % daily error with collinearity is significant for MV and Sulfur. For MV, it increases from 21 to 31%, while for S, it increases from 25 to 29% as collinearity increases from 0 to 1.2. The % error for the estimation of average mass contributions is 23, 80, and 23% for MV, NaCl, and S, respectively. On average, the “Sodium Chloride” contribution is overestimated, while “Motor Vehicle” and “Sulfur” contributions are underestimated. As collinearity increases from 0 to 1.2, the % error for the estimation of average mass contributions significantly increases

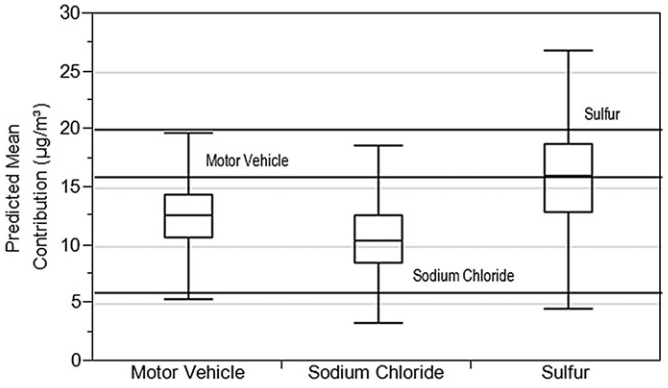

Fig. 1. Mean predicted mass contributions (µg m^{-3}) for the three sources across all collinearity scenarios in the primary analysis (n = 2700 datasets). The horizontal reference lines indicate the simulated (input) mean mass contributions.

from 17 to 27% for MV, from 77 to 93% for NaCl, and from 21 to 25% for S.

The average Pearson correlation coefficient between the simulated and predicted daily factor contributions (*G* correlation) decreases significantly as collinearity increases in all three simulated factors in the primary analysis (Fig. 2A). The sharpest decrease is seen in "Motor Vehicle", followed by "Sulfur" and then by "Sodium Chloride", following the same ordering of the assumed CV's for the three sources. Finally, the *F* correlation, or the correlation between the simulated and predicted factor profiles, also decreases significantly as collinearity increases (Fig. 3A).

3.2. Sensitivity analyses

Table 4 shows the % absolute errors for estimating daily and average source contributions in the different simulation scenarios. As previously stated, the % daily and average errors increase with collinearity. In the two scenarios where the mean input mass contribution of NaCl is increased from 6 to 10 µg m^{-3} and from 6 to 16 µg m^{-3}, the % absolute error for average source contribution decreases from 80 to 33 and 18%, respectively. Increasing the variance of MV decreases its daily and average estimation errors. Setting the mean mass contribution of the three sources to 16 µg m^{-3} results in the lowest overall daily and average errors. When the CV's of the three sources are set to the same level, the NaCl source exhibits the largest prediction errors, followed by S then MV.

Fig. 2 shows the *G* correlations, or correlations between the simulated and predicted source contributions, against collinearity in each simulation. In the primary analysis, the ability of PMF to predict the source contributions decreases significantly as collinearity increases for all factors. "Sodium Chloride" with the highest CV in its source contribution performs the best, while "Motor Vehicle" with the lowest CV performs the worst. Moreover, the CV of the source contributions influences the PMF solution in the presence of collinearity. When the variance of NaCl increases from 6 to 16, and that of MV increases from 21 to 40, the deterioration in the *G* correlations with increasing collinearity becomes non-significant. In contrast, decreasing the variance of NaCl from 6 to 3 results in a greater and more significant deterioration in the PMF solution, in terms of being able to predict the source contributions. By increasing the CV of all three source contributions to the same level of 0.4, the decrease in *G* correlations becomes smaller but still significant.

Keeping the CV constant, as the mean contribution of NaCl increases from 6 to 10 µg m^{-3}, the decrease in *G* correlations against increasing collinearity becomes non-significant. Setting the mean contributions of all three sources to 16 µg m^{-3}, the decrease in *G* correlations as collinearity increases is non-significant for NaCl, and it is significant for MV and S. In this simulation, NaCl is the only source whose mass is increased compared to the primary analysis. Finally, varying the input factor profiles to become more similar and share more elements results in non-significant decreases in the *G* correlations with increasing collinearity for all three sources.

Fig. 3 shows the average Pearson correlation coefficients between the simulated and predicted profiles (*F* correlations) by collinearity for each source and simulation. In the primary analysis, the ability of PMF to predict the source profiles deteriorates significantly as collinearity in the contributions of the three sources increases. The "Sulfur" factor with the highest mean contribution seems to perform best. PMF is more robust to high collinearity when the mean contribution of NaCl is increased, evidenced by the smaller but still significant decrease in *F* correlations of NaCl compared to the primary analysis.

By separately increasing the variance of the NaCl or MV source contributions, the decrease in *F* correlations becomes smaller and less significant for all three factors, suggesting an improved performance of PMF in the presence of increasing collinearity. Conversely, decreasing the NaCl variance results in a sharper and significant decrease in *F* correlations. When the CV of all factors increases to 0.4, keeping the mean contributions constant, the performance of PMF in the face of increasing collinearity improves for all three sources, even though the decrease in *F* correlations remains significant. "Sulfur" with the highest mean contribution performs best.

Assigning the same mean contribution of 16 µg m^{-3} to all factors while keeping their original CV's also results in significant decreases in the *F* correlation with collinearity. However, the decrease in *F* correlations becomes smaller for NaCl and larger for S. This is reflecting the change in the contributions compared to the primary analysis, since the NaCl mean contribution increases from 6 to 16 µg m^{-3} and that of S decreases from 20 to 16 µg m^{-3}. Finally, in the similar profiles scenario, the *F* correlation does not show any significant change with increasing collinearity.

4. Discussion

We simulated fine particle and elemental mass concentrations and their uncertainties based on characteristics of real data to make this simulation exercise as realistic as possible. We created 27 basic scenarios that correspond to different degrees of collinearity amongst the contributions of three sources, and simulated these scenarios 100 times. Subsequently, we varied the mean source contributions, variances of the source contributions, and profiles of these sources in seven different sensitivity analysis scenarios, and examined the impact of each scenario on the PMF solution stability while increasing source collinearity. The rationale for our selection of the seven sensitivity analyses was to first confirm our suspicion that the CV of a factor is highly influential in determining how well it is predicted by PMF in the presence of collinearity, and secondly, to further investigate reports from the literature suggesting that factors with higher mass contributions or with unique profiles are predicted with lower error.

We first targeted "Sodium Chloride", the factor that PMF resolved with the least error in the primary analysis in the presence of increasing collinearity. We changed the mean and variance of its contribution in order to determine the influence of these two parameters on the robustness of the PMF solution. We also changed the variance of NaCl to 3, 8, 10 and 16, but only presented the lowest and highest variance scenarios since the remaining results followed the expected direction. We then changed the

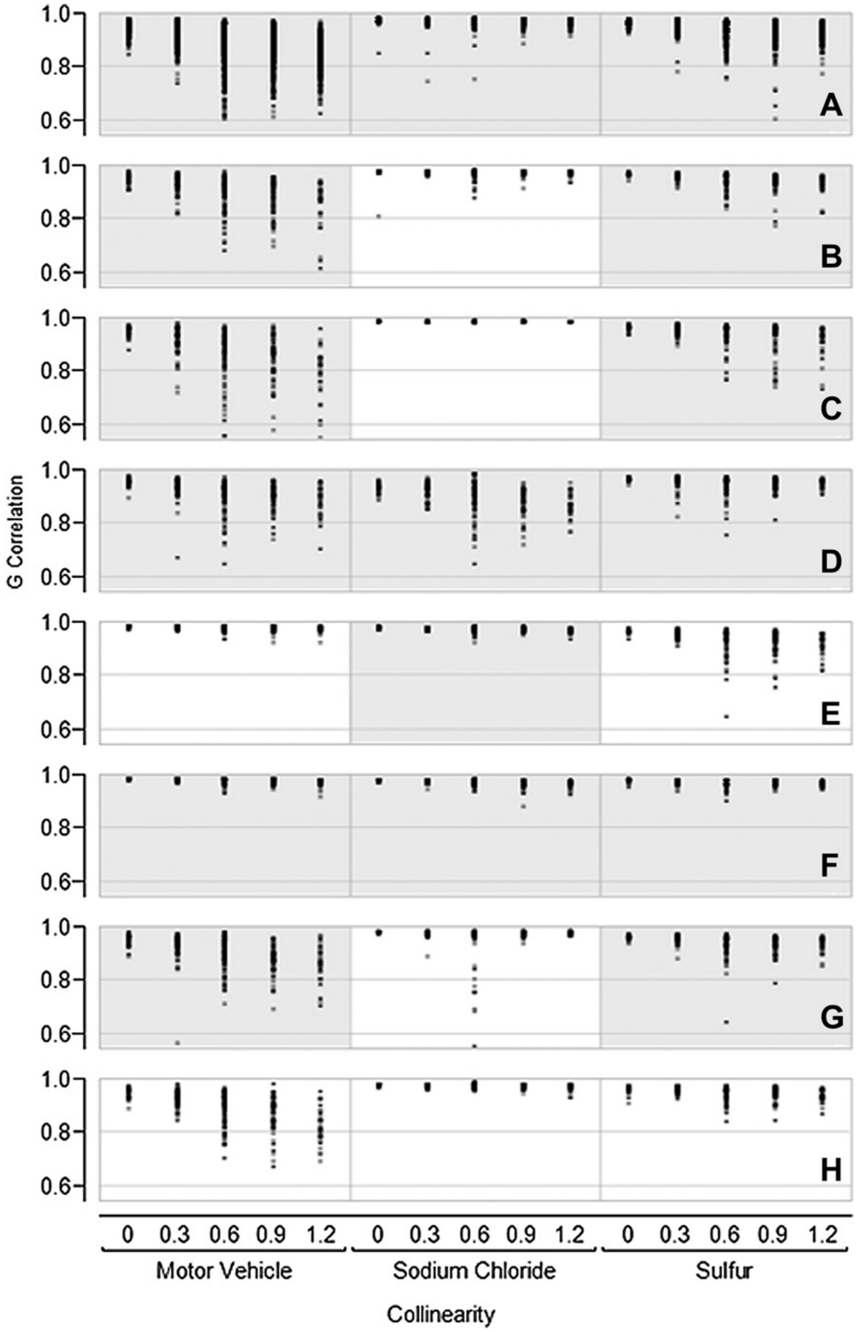

Fig. 2. *G* correlation or Pearson correlation coefficient between simulated and predicted daily mass contributions against collinearity for each source in the primary analysis (A) and the seven sensitivity analyses: increase NaCl mean (B), increase NaCl variance (C), decrease NaCl variance (D), increase MV variance (E), same CVs (F), same means (G), and similar profiles (H). Simulations where the decrease in correlation is significant at the 0.05 level between the observed and predicted *G* as a function of source collinearity are shaded.

variance of "Motor Vehicle" to confirm the findings of NaCl using a different factor. Following that, we set the CV's to the same level to look at the independent effect of the mean contributions, and then set the mean contributions to the same level to look at the effect of the CV's independently. We finally made the profiles less distinguishable.

All datasets were analyzed in a similar manner in order to isolate the effect of the varying correlation structures between the source contributions. In order to simulate normally-distributed and positive source contributions, we had to create concentrations that are normally-distributed and non-negative. As a result the generated mass and elemental concentrations were higher than those typically observed. However, the PMF performance was stable across the different simulations. For example, in the primary analysis of 2700 datasets, the Q_{robust} goodness-of-fit parameter values were stable and ranged between 2365 and 2686. The PMF solutions were similarly stable in the seven sensitivity analyses. This confirmed our assumption that the selected close-to-default PMF settings for our simulations were suitable. Moreover, Lingwall and Christensen (2007) tested several PMF settings in their simulation study that looked at the use of *F* element pulling or profile targeting in PMF. They found that the default settings are generally appropriate, and very few of them had a dramatic effect on the model performance.

In our simulations, source collinearity was not as influential as the mass contribution in determining the level of error in

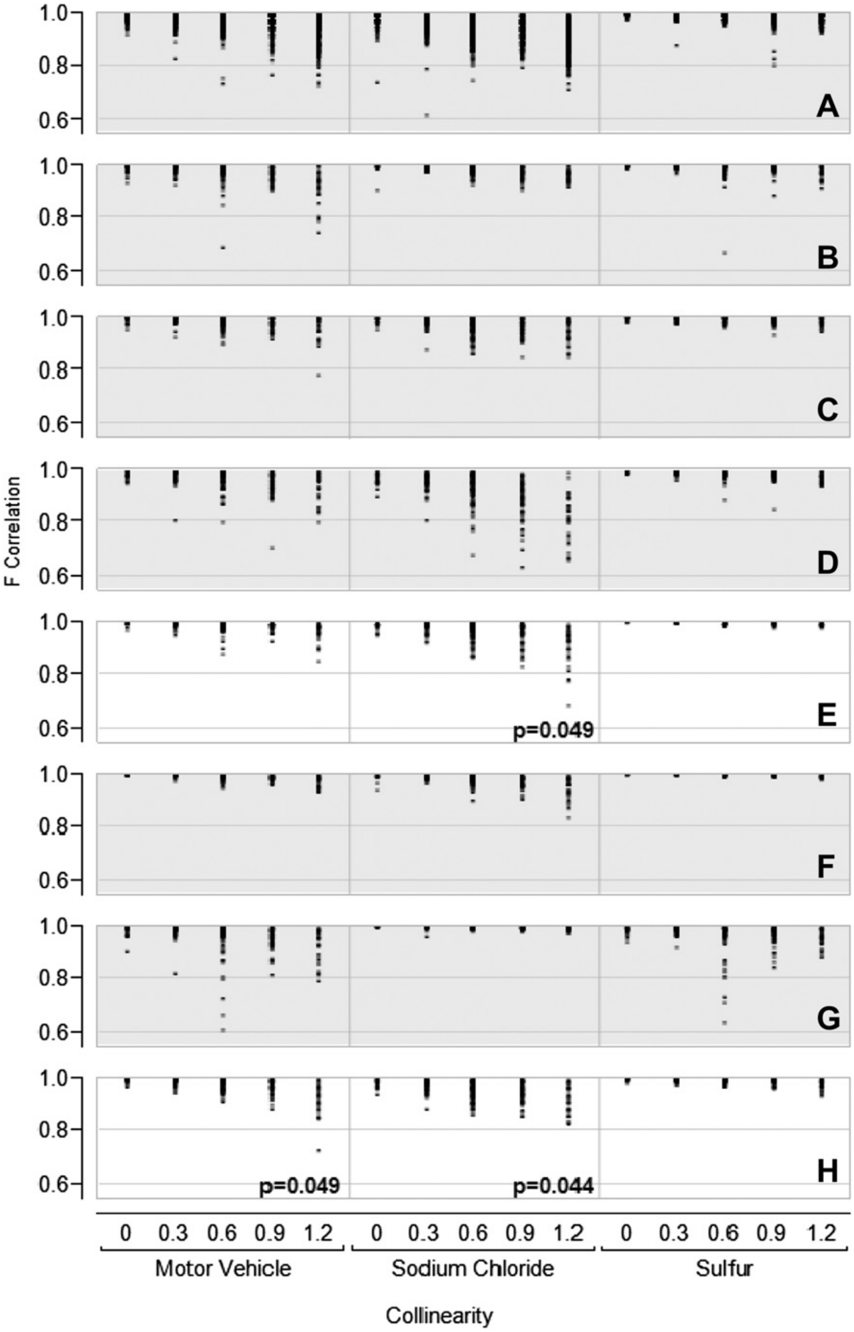

Fig. 3. *F* correlation or Pearson correlation coefficient between simulated and predicted source profiles against collinearity for each source in the primary analysis (A) and the seven sensitivity analyses: increase NaCl mean (B), increase NaCl variance (C), decrease NaCl variance (D), increase MV variance (E), same CVs (F), same means (G), and similar profiles (H). Simulations where the decrease in correlation is significant at the 0.05 level between the observed and predicted *F* as a function of source collinearity are shaded. The *p*-value is indicated for scenarios where the decrease in *F* correlation is marginally significant.

estimating daily and average source contributions. For example, the % average error for NaCl increased from 77 to 93% as collinearity increased from 0 to 1.2 in the primary analysis. Whereas, when the NaCl mass contribution was increased from 6 to 10 μg m^{-3}, its % average error decreased from a mean of 80 to 33%.

In all the scenarios where the mean simulated mass contributions of MV, NaCl, and S were 16, 6 and 20 μg m^{-3}, respectively, PMF consistently underpredicted the MV and S average contributions and overpredicted the NaCl contribution. This could be due to the fact that PMF tends to apportion the daily $PM_{2.5}$ mass somewhat equally among the three sources, resulting in predicted contributions that are closer to the overall mean (14 μg m^{-3}) of the input source contributions. This is further evidenced in the “Same Means” scenario, where the mean mass contributions were all set to 16 μg m^{-3}. Despite the fact that the mean mass of NaCl increased from 6 to 16 μg m^{-3} but that of S decreased from 20 to 16 μg m^{-3}, the % average errors decreased from 23, 80 and 23% in the primary analysis, to 19, 18 and 14% for MV, NaCl, and S, respectively. This improvement in the PMF solution is likely due to decreasing, or in this case, eliminating the difference in the input contributions of the simulated sources.

Furthermore, we found that the *G* correlations decreased significantly as collinearity in the source contributions increased,

Table 4
Mean absolute errors (%) for estimating daily and average source contributions in the primary analysis and sensitivity analyses.

Scenarios	Absolute Errors (%)					
	Motor Vehicle		Sodium Chloride		Sulfur	
	Daily	Average	Daily	Average	Daily	Average
Primary Analysis	27	23	83	80	27	23
Sensitivity Analyses						
1) Increase NaCl Mean	29	20	38	33	23	19
2) Increase NaCl Variance	27	22	88	78	26	21
3) Decrease NaCl Variance	31	27	59	57	28	24
4) Increase MV Variance	22	16	92	89	26	22
5) Same Coefficients of Variation	20	15	83	75	23	20
6) Same Means	24	19	23	18	19	14
7) Similar Profiles	26	22	79	77	27	24

implying that the PMF solution deteriorated. As collinearity increased, the factor with the lowest CV, MV in the primary analysis, performed the worst, while the factor with the highest CV, NaCl, performed the best. The importance of the CV was confirmed by increasing the CV of NaCl and MV separately, in two sensitivity analyses, and finding that the decrease in *G* correlations with increasing collinearity became smaller and non-significant. On the contrary, when the CV of NaCl was decreased, the deterioration of the *G* correlations became more pronounced and significant. Similarly, increasing the CV of all three factors improved PMF performance against increasing collinearity.

We also found that increasing a factor's mean contribution resulted in an improved performance against collinearity. When the mean of NaCl was increased, the *G* correlations still decreased with increasing collinearity; however, this decrease was non-significant. Setting the same CV's, the *G* correlations decreased significantly with increasing collinearity in all three sources, despite their differing mean mass contributions. Finally, in the "Similar Profiles" scenario, when the three factor profiles were simulated to share more elements, the effect of collinearity on *G* correlations became smaller and non-significant, and NaCl, with the highest CV among the three factors, was predicted with the least error.

As for the factor profiles, the ability of PMF to predict them in the face of increasing collinearity is reflected in the *F* correlations, which are the Pearson correlation coefficients between the simulated and the predicted factor loading profiles. We found that the *F* correlations decreased significantly in all three sources with increasing collinearity. However, the source with the highest mean contribution, "Sulfur", performed the best despite not having the highest CV. Similarly, increasing the mean contribution of a factor improved PMF's ability to predict its loading profile as collinearity increased. Also, increasing a factor's CV rendered the effect of collinearity on *F* correlations smaller and less significant. In the scenario where factor profiles were simulated to be more similar, the decrease in *F* correlations was non-significant. PMF predicted the factor profiles with a similar degree of error across all collinearity levels.

Our findings agree with those reported by other simulation studies which found that correlated sources were harder to separate and their predictions encompassed higher error and bias. Miller et al. (2002) found that even though PMF performed best in extracting the major source profiles of their simulated volatile organic compounds (VOC's) personal exposure sources, it had difficulty extracting sources that had similar chemical profiles. In addition, all four receptor models evaluated, including PMF, CMB, PCA/APCS and UNMIX, were sensitive to collinearities in the simulated data induced by common meteorological effects on VOC's source contributions. Also, all models had difficulty identifying sources that contributed less than 5% to the total VOC concentrations. Christensen and Schauer (2008) found that source contribution errors were relatively lower for sources encompassing the secondary species sulfate and nitrate that are usually present at high concentrations and measured with the least error. Contribution estimates of the sulfate and nitrate related sources were more stable against perturbations in the uncertainties matrix specified in PMF.

In the Brinkman et al. (2006) simulations, the "vegetative debris" factor that contributed about 5% to $PM_{2.5}$ mass and was the least variable was not properly identified. Also factors that had highly correlated simulated mass contributions were not well identified. Furthermore, Hemann et al. (2009) found that in a simulated dataset the three factors that had moderately strong correlations with each other had poor solutions with respect to bias, variability, and temporal pattern prediction.

Our simulations could have benefitted from a greater range of correlations and sensitivity analyses, but we were limited by timing and computing constraints. We could have also attempted rotating the PMF predicted factor profiles to match the input profiles; however, we chose to treat the scenarios identically to differentiate the impact of collinearity, and use settings that are as close to default as possible in PMF. We could also compare the performance of PMF to other source apportionment models like UNMIX or PCA. Furthermore, our measure of collinearity is an approximation of the correlation affecting each source and it may not be ideal to differentiate the impact of correlation between each pair of factors.

The main focus of our analysis was to systematically investigate the impact of source collinearity on the PMF receptor model solution. The use of simulated data makes it possible to compare model inputs to outputs for different collinearity scenarios, while controlling for other sources of variability that may exist in real scenarios, such as differences in the number of contributing sources, changes in their day-to-day profiles, and other factors. However, our findings depend on the characteristics of the simulated data used for the PMF analysis and may not reflect its performance when used to analyze real particle concentration data.

Sources with higher coefficients of variation, as represented by tracer species that have higher coefficients of variation, were better predicted than sources with lower CV's when collinearity existed. Therefore, this exercise can be helpful in identifying situations where source collinearity might exist, and may warrant combining receptor modeling predictions with further confirmatory analyses and physical parameters or supporting information, like wind trajectory analyses, or including more tracer species or pollutants in the analyses. In such situations, one should be cautious when interpreting model predictions and using estimated source contributions in health effect analyses. As the correlation between the actual and predicted source contributions of certain sources degrades, the power to detect significant concentration-response estimates weakens, biasing toward the null. This might also lead to biased health effect risk estimates or to attributing the impact of one source group to another. As reported by Ito et al. (2004), local sources might contain higher degrees of error than regional sources and secondary aerosols in source apportionment studies. In such instances, the use of source apportioned particulate matter contributions in morbidity and mortality health effect analyses may result in distorted or undetected estimates of absolute health risk per unit mass concentration.

Acknowledgments

This publication was made possible by USEPA grant RD 83479801. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

Brinkman, G., Vance, G., Hannigan, M.P., Milford, J.B., 2006. Use of synthetic data to evaluate positive matrix factorization as a source apportionment tool for $PM_{2.5}$ exposure data. Environmental Science & Technology 40, 1892–1901.

Bzdusek, P.A., Christensen, E.R., 2006. Comparison of a new variant of PMF with other receptor modeling methods using artificial and real sediment PCB data sets. Environmetrics 17, 387–403.

Christensen, W.F., Schauer, J.J., 2008. Impact of species uncertainty perturbation on the solution stability of positive matrix factorization of atmospheric particulate matter data. Environmental Science & Technology 42, 6015–6021.

Dockery, D.W., Pope, C.A., Xu, X.P., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Speizer, F.E., 1993. An association between air-pollution and mortality in 6 United-States cities. New England Journal of Medicine 329, 1753–1759.

Dutton, S.J., Schauer, J.J., Vedal, S., Hannigan, M.P., 2009. $PM_{2.5}$ characterization for time series studies: pointwise uncertainty estimation and bulk speciation methods applied in Denver. Atmospheric Environment 43, 1136–1146.

Gent, J.F., Koutrakis, P., Belanger, K., Triche, E., Holford, T.R., Bracken, M.B., Leaderer, B.P., 2009. Symptoms and medication use in children with asthma and traffic-related sources of fine particle pollution. Environmental Health Perspectives 117, 1168–1174.

Grahame, T., Hidy, G.M., 2007. Pinnacles and pitfalls for source apportionment of potential health effects from airborne particle exposure. Inhalation Toxicology 19, 727–744.

Hemann, J.G., Brinkman, G.L., Dutton, S.J., Hannigan, M.P., Milford, J.B., Miller, S.L., 2009. Assessing positive matrix factorization model fit: a new method to estimate uncertainty and bias in factor contributions at the measurement time scale. Atmospheric Chemistry and Physics 9, 497–513.

Hopke, P.K., Ito, K., Mar, T., Christensen, W.F., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Larson, T.V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., Paatero, P., Thurston, G.D., 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. Journal of Exposure Science and Environmental Epidemiology 16, 275–286.

Ito, K., Christensen, W.F., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Larson, T.V., Neas, L., Hopke, P.K., Thurston, G.D., 2006. PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. Journal of Exposure Science and Environmental Epidemiology 16, 300–310.

Ito, K., Xue, N., Thurston, G., 2004. Spatial variation of $PM_{2.5}$ chemical species and source-apportioned mass concentrations in New York City. Atmospheric Environment 38, 5269–5282.

Kim, E., Hopke, P.K., 2007. Comparison between sample-species specific uncertainties and estimated uncertainties for the source apportionment of the speciation trends network data. Atmospheric Environment 41, 567–575.

Laden, F., Neas, L.M., Dockery, D.W., Schwartz, J., 2000. Association of fine particulate matter from different sources with daily mortality in six US cities. Environmental Health Perspectives 108, 941–947.

Larsen, R.K., Baker, J.E., 2003. Source apportionment of polycyclic aromatic hydrocarbons in the urban atmosphere: a comparison of three methods. Environmental Science & Technology 37, 1873–1881.

Lee, E., Chan, C.K., Paatero, P., 1999. Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. Atmospheric Environment 33, 3201–3212.

Lee, J.H., Hopke, P.K., Turner, J.R., 2006. Source identification of airborne $PM_{2.5}$ at the St. Louis-Midwest Supersite. Journal of Geophysical Research-Atmospheres 111, 12.

Lingwall, J.W., Christensen, W.F., 2007. Pollution source apportionment using a priori information and positive matrix factorization. Chemometrics and Intelligent Laboratory Systems 87, 281–294.

Miller, S.L., Anderson, M.J., Daly, E.P., Milford, J.B., 2002. Source apportionment of exposures to volatile organic compounds. I. Evaluation of receptor models using simulated exposure data. Atmospheric Environment 36, 3629–3641.

Norris, G., Vedantham, R., Wade, K., Brown, S., Prouty, J., Foley, C., 2008. EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide.

Norris, G., Vedantham, R., Wade, K., Zahn, P., Brown, S., Paatero, P., Eberly, S., Foley, C., 2009. Guidance Document for PMF Applications with the Multilinear Engine.

Paatero, P., 1999. The multilinear engine - a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. Journal of Computational and Graphical Statistics 8, 854–888.

Paatero, P., 2010. User's Guide for Positive Matrix Factorization Programs PMF2 and PMF3, Parts 1 and 2.

Paatero, P., Hopke, P.K., 2003. Discarding or downweighting high-noise variables in factor analytic models. Analytica Chimica Acta 490, 277–289.

Paatero, P., Hopke, P.K., Begum, B.A., Biswas, S.K., 2005. A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution. Atmospheric Environment 39, 193–201.

Paatero, P., Tapper, U., 1993. Analysis of different modes of factor-analysis as least-squares fit problems. Chemometrics and Intelligent Laboratory Systems 18, 183–194.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization - a nonnegative factor model with optimal utilization of error-estimates of data values. Environmetrics 5, 111–126.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama-Journal of the American Medical Association 287, 1132–1141.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing Vienna, Austria.

Reff, A., Eberly, S.I., Bhave, P.V., 2007. Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods. Journal of the Air & Waste Management Association 57, 146–154.

SAS Institute Inc., 2011. SAS 9.2 Cary, NC, USA.

Thurston, G., Ito, K., Mar, T., Christensen, W.F., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Larson, T.V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., Hopke, P.K., 2005. Results and implications of the workshop on the source apportionment of PM health effects. Epidemiology 16, S134–S135.